(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 115 225 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
11.07.2001 Bulletin 2001/28

(51) Int Cl.[7]: **H04L 12/24**

(21) Application number: **00311661.3**

(22) Date of filing: **22.12.2000**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **06.01.2000 US 478306**

(71) Applicant: **International Business Machines Corporation**
Armonk, NY 10504 (US)

(72) Inventors:
- **Barnett, Barry Stanley, IBM United Kingdom Ltd.**
  Winchester, Hampshire SO21 2JN (GB)
- **Bossen, Douglas Craig,**
  IBM United Kingdom Ltd.
  Winchester, Hampshire SO21 2JN (GB)

(74) Representative: **Jennings, Michael John**
IBM United Kingdom Limited,
Intellectual Property Department,
Hursley Park
Winchester, Hampshire SO21 2JN (GB)

(54) Method and system for end-to-end problem determination and fault isolation for storage area networks

(57) A method and system for problem determination and fault isolation in a storage area network (SAN) is provided. A complex configuration of multi-vendor host systems, FC switches, and storage peripherals are connected in a SAN via a communications architecture (CA). A communications architecture element (CAE) is a network-connected device that has successfully registered with a communications architecture manager (CAM) on a host computer via a network service protocol, and the CAM contains problem determination (PD) functionality for the SAN and maintains a SAN PD information table (SPDIT). The CA comprises all network-connected elements capable of communicating information stored in the SPDIT. The CAM uses a SAN topology map and the SPDIT are used to create a SAN diagnostic table (SDT). A failing component in a particular device may generate errors that cause devices along the same network connection path to generate errors. As the CAM receives error packets or error messages, the errors are stored in the SDT, and each error is analyzed by temporally and spatially comparing the error with other errors in the SDT. If a CAE is determined to be a candidate for generating the error, then the CAE is reported for replacement if possible.

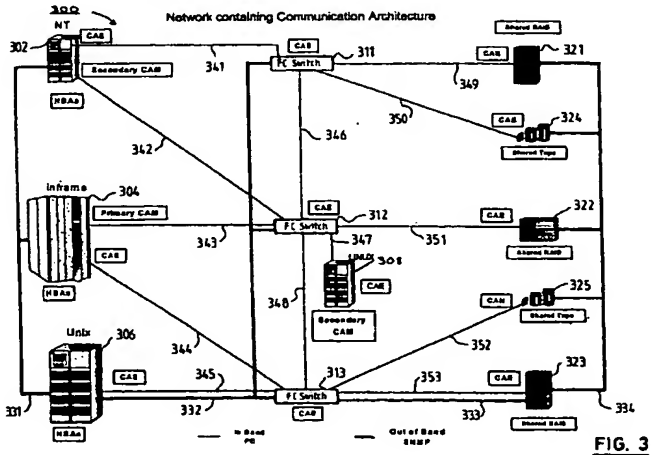

FIG. 3

EP 1 115 225 A2

Description

[0001] The present invention relates to an improved data processing system and, in particular, to a method and apparatus for computer network managing.

[0002] A Storage Area Network (SAN) is an "open system" storage architecture that allows multiple host computers to share multiple storage peripherals, and in particular, to share storage peripherals via a Fibre Channel (FC) network switch. The FC switch, host systems, and storage peripherals may be manufactured by different vendors and contain different operating environments.

[0003] Currently, there is a lack of an end-to-end problem determination capability or specification for an FC SAN. A complex configuration of multi-vendor systems, network switches, and peripherals makes it significantly more difficult to perform problem determination in a SAN environment than existing point-to-point storage configurations. As a result, failures in a SAN environment will cause an increase of system downtime as well as increasing cost of system maintenance.

[0004] Accordingly, the invention provides a method for processing errors within a storage area network (SAN), the method comprising the computer-implemented steps of: generating a SAN topology map; generating a SAN problem determination information table (SPDIT); and generating a SAN diagnostic table (SDT) using the SAN topology map and the SPDIT.

[0005] The invention further provides a data processing system for communicating error information in a storage area network (SAN), the data processing system comprising: a network comprising in-band Fibre Channel communication links and out-of-band communication links, wherein the network supports a communications architecture (CA); a plurality of storage devices connected to the network; a plurality of host computers connected to the network, wherein at least one of the plurality of host computers comprises a communications architecture manager (CAM) containing problem determination (PD) functionality, wherein a CAM maintains a SAN PD information table (SPDIT), and wherein the CA comprises all network-connected elements capable of communicating information stored in the SPDIT.

[0006] The invention still further provides a data processing system for processing errors within a storage area network (SAN), the data processing system comprising: first generating means for generating a SAN topology map; second generating means for generating a SAN problem determination information table (SPDIT); and third generating means for generating a SAN diagnostic table (SDT) using the SAN topology map and the SPDIT.

[0007] The invention still further provides a computer program comprising program code for controlling a data processing system to perform operations implementing a method for processing errors within a storage area network (SAN), the program code comprising: first instructions for generating a SAN topology map; second instructions for generating a SAN problem. determination information table (SPDIT); and third instructions for generating a SAN diagnostic table (SDT) using the SAN topology map and the SPDIT.

[0008] According to a preferred embodiment, a method and system for problem determination and fault isolation in a storage area network (SAN) is provided.

[0009] Preferably, a method and apparatus are provided that define an "open system", real-time, end-to-end, error detection architecture that incorporates fault isolation algorithms to identify failing systems and/or components connected to a SAN.

[0010] According to a preferred embodiment, a complex configuration of multi-vendor host systems, FC switches, and storage peripherals are connected in a SAN via a communications architecture (CA). A communications architecture element (CAE) is a network-connected device that has successfully registered with a communications architecture manager (CAM) on a host computer via a network service protocol, and the CAM contains problem determination (PD) functionality for the SAN and maintains a SAN PD information table (SPDIT).

[0011] The CA preferably comprises all network-connected elements capable of communicating information stored in the SPDIT. The CAM uses a SAN topology map and the SPDIT to create a SAN diagnostic table (SDT). A failing component in a particular device may generate errors that cause devices along the same network connection path to generate errors. As the CAM receives error packets or error messages, the errors are preferably stored in the SDT, and each error is analyzed by temporally and spatially comparing the error with other errors in the SDT. If a CAE is determined to be a candidate for generating the error, then the CAE is reported for replacement if possible.

[0012] A preferred embodiment of the present invention will now be described, by way of example only, with reference to the following drawings:

Figure 1 is a pictorial representation depicting a data processing system in which the preferred embodiment of the present invention is implemented;

Figure 2 is an example block diagram illustrating internal components of a server-type data processing system that implements the present invention according to a preferred embodiment;

Figure 3 is a diagram depicting a communications architecture for data processing systems that participate in the SAN problem determination methodology implemented in accordance with a preferred embodiment of the present invention;

Figure 4 is a table depicting a SAN Problem Deter-

mination Information Table (SPDIT) in accordance with a preferred embodiment of the present invention;

Figure 5A is a simplified network topology diagram for a SAN in accordance with a preferred embodiment;

Figure 5B is a table providing a topology map for the SAN shown in **Figure 5A** in accordance with a preferred embodiment of the present invention;

Figure 6 is a diagram depicting a SAN Diagnostic Table for a SAN in accordance with a preferred embodiment of the present invention;

Figure 7 is a table depicting the weightings to be used in real-time diagnostic analysis for various errors in accordance with a preferred embodiment of the present invention; and

Figures 8A-8D are flowcharts depicting a process for a real-time diagnostic algorithm for SAN end-to-end fault isolation of a single failing SAN element in accordance with a preferred embodiment of the present invention.

[0013] With reference now to **Figure 1**, a pictorial representation depicts a data processing system in which a preferred embodiment of the present invention is implemented. A computer **100** is depicted, which includes a system unit **110**, a video display terminal **102**, a keyboard **104**, storage devices **108**, which may include floppy drives and other types of permanent and removable storage media, and mouse **106**. Additional input devices may be included with computer **100**. Computer **100** can be implemented using any suitable computer, for example, an IBM RISC/System 6000 system, a product of International Business Machines Corporation in Armonk, New York, running the Advanced Interactive Executive (AIX) operating system, also a product of IBM Corporation. Although the depicted representation shows a server-type computer, other embodiments of the present invention are implemented in other types of data processing systems, such as workstations, network computers, Web-based television set-top boxes, Internet appliances, etc. Computer **100** also preferably includes a graphical user interface that is implemented by means of system software residing in computer readable media in operation within computer **100**.

[0014] **Figure 1** is intended as an example and not as an architectural limitation for the present invention.

[0015] With reference now to **Figure 2**, a block diagram depicts a typical organization of internal components in a data processing system. Data processing system **200** employs a variety of bus structures and protocols. Although the depicted example employs a PCI bus, an ISA bus, and a 6XX bus, other bus architectures and protocols may be used.

[0016] Processor card **201** contains processor **202** and L2 cache **203** that are connected to 6XX bus **205**. System **200** may contain a plurality of processor cards. Processor card **206** contains processor **207** and L2 cache **208**.

[0017] 6XX bus **205** supports system planar **210** that contains 6XX bridge **211** and memory controller **212** that supports memory card **213**. Memory card **213** contains local memory **214** consisting of a plurality of dual in-line memory modules (DIMMs) **215** and **216**.

[0018] 6XX bridge **211** connects to PCI bridges **220** and **221** via system bus **222**. PCI bridges **220** and **221** are contained on native I/O (NIO) planar **223** which supports a variety of I/O components and interfaces. PCI bridge **221** provides connections for external data streams through network adapter **224** and a number of card slots **225-226** via PCI bus **227**. PCI bridge **220** connects a variety of I/O devices via PCI bus **228**. Hard disk **229** may be connected to SCSI host adapter **230**, which is connected to PCI bus **228**. Graphics adapter **231** may also be connected to PCI bus **228** as depicted, either directly or indirectly.

[0019] ISA bridge **232** connects to PCI bridge **220** via PCI bus **228**. ISA bridge **232** provides interconnection capabilities through NIO controller **233** via ISA bus **234**, such as serial connections **235** and **236**. Floppy drive connection **237** provides removable storage. Keyboard connection **238** and mouse connection **239** allow data processing system **200** to accept input data from a user. Non-volatile RAM (NVRAM) **240** provides non-volatile memory for preserving certain types of data from system disruptions or system failures, such as power supply problems. System firmware **241** is also connected to ISA bus **234** and controls the initial BIOS. Service processor **244** is connected to ISA bus **234** and provides functionality for system diagnostics or system servicing.

[0020] Service processor **244** detects errors and passes information to the operating system. The source of the errors may or may not be known to a reasonable certainty at the time that the error is detected. The operating system may merely log the errors or may otherwise process reported errors.

[0021] Those of ordinary skill in the art will appreciate that the hardware in **Figure 2** may vary depending on the system implementation. For example, the system may have more processors, and other peripheral devices may be used in addition to or in place of the hardware depicted in **Figure 2**. The depicted examples are not meant to imply architectural limitations with respect to the present invention.

[0022] With reference now to **Figure 3**, a diagram depicts a communications architecture for data processing systems that participate in the SAN problem determination methodology implemented in accordance with a preferred embodiment of the present invention. Network 300 comprises a set of computers, switches, and storage devices that may or may not participate in the com-

munications architectures.

[0023] The Communications Architecture (CA) comprises all SAN-connected elements capable of communicating any or all of the information defined in a SAN Problem Determination Information Table (SPDIT), which is described in more detail further below.

[0024] Each SAN connected element participating in the CA is called a CA Element (CAE). Any element not participating in the CA is called a CA Non-participant (CAN). These elements are distinguished because they both participate in the SAN topology and thereby the problem determination (PD) capabilities of the system. Windows NT™ server **302**, mainframe computer **304**, Unix™ server **306**, and Linux™ server **308** are computers that participate in the CA and are thus CAEs. Windows NT™ server **302**, mainframe computer **304**, and Unix™ server **306** are also host computers that may support various clients, which may require access to the storage devices. Each of computers **302-306** has a Host Bus Attach (HBA), which is a type of network adapter for FC hosts. FC switches **311-313** are CAEs, and some of the storage devices are also CAEs. In the example, shared RAIDS (Redundant Array of Independent Disks) **321-323** and shared tape **324** are CAEs, while shared tape **325** is a CAN.

[0025] The CA can communicate via the FC switching fabric via in-band communication links **341-352** using the TCP/IP protocol and/or via an out-of-band TCP/IP communication network on communication links **331-334** that all SAN elements share. It should be noted that the communication links depicted in **Figure 3** may be logical connections that share a single physical connection. Alternatively, the devices may be connected by more than one physical communication link.

[0026] The protocols used by the CA to issue and/or collect information are defined to be both SNMP/MIB (Simple Network Management Protocol/Management Information Base, an SNMP structure that describes the particular device being monitored) and native FC based. The use of these two protocols allows both device/host-specific and SAN-specific information to be collected and subsequently used for end-to-end problem determination.

[0027] CA Managers (CAMs) are special CAEs in which the end-to-end PD capabilities of the system reside. The SPDIT resides in the CAM and every CAE is automatically registered with a CAM (via native FC and/or SNMP services). CAEs are those elements that successfully register, and CANs are those elements that cannot register with the CAM but are known to be present via the SAN topology discovery process, which is discussed in more detail further below. CAMS support any FC Extended Link Services (ELS) that are relevant to end-to-end problem determination.

[0028] CAMS may be categorized as a primary or active CAM and secondary or inactive CAMs. CAMS are highly available elements that replicate SPDIT and registration information. For example, a secondary CAM and a primary CAM may share a heartbeat signal so that a secondary CAM, operating in a redundant manner, may assume the duties of the primary CAM if the primary CAM appears to have failed by not responding to the heartbeat signal. The problem determination interface to the CAM is comprised of a SAN PD Application Programming Interface (SAN PD API). The SAN PD API defines the communication interface between the CAM and any other operating environment that can read CAM information or status.

[0029] With reference now to **Figure 4**, a table depicts a SAN Problem Determination Information Table (SPDIT) in accordance with a preferred embodiment of the present invention. The SPDIT is comprised of all known products/elements and the information types that can be communicated on the CA. The format of the SPDIT may vary depending upon the number of devices in the CA, the type of products that are supported, the information associated with the devices, etc. For example, the SPDIT would contain information concerning each device shown in **Figure 3**.

[0030] SPDIT **400** contains, by way of example, the following record entries:

vendor attribute **401**, product identifier **402**, info type **403**, and description **404**. Each record in SPDIT **400** contains data for these record entries. Vendor attribute **401** contains the manufacturer of a particular device on the CA. Product identifier **402** contains vendor-assigned information for identifying a particular device, such as model type, model number, product serial number, etc.

[0031] Information type **403** contains data related to the type of communication links supported by the device, the format of error conditions or error definitions supported by the device, etc. Description attribute **404** provides information about the type of error information that should be expected by the product. For example, if the description attribute record only contains an indication that the product is ELS Registered Link Incident Record (RLIR) compatible, then a CAM-related process would not expect to receive out-of-band MIBs for the product.

[0032] The SPDIT will generally contain all information used to indicate status/error conditions by SAN capable peripherals, hosts, and switches. This would include native FC link and extended link error definitions, and MIB definitions. These definitions can include field replaceable unit (FRU) component information, which can be located in a MIB or embedded in the error reporting protocol and can be used to determine the granularity to which failing components can be isolated.

[0033] As noted previously, CAMs are special CAs in which the end-to-end PD capabilities of the system reside, including the SPDIT. The CAM initialization process includes the discovery and registration of all FC nodes connected to both the in-band SAN and out-of-band network. The CAM initialization process uses FC in-band and CA out-of-band (via SNMP) discovery/registration processes. This process provides a topology

map (TM) of the SAN that includes all registered and non-registered SAN connected elements along with knowledge of the element types (hosts, peripherals, switches), explicit connections/paths, and their relevant vendor and SPDIT information.

[0034] With reference now to **Figure 5A**, a simplified network topology diagram for the SAN in accordance with a preferred embodiment is shown. FC switch **501** contains ports **511-513** providing connection points between FC switch **501** and CAEs **521-523**, also labelled CAE A, CAE B, and CAE C. From the perspective of the CA, FC switch ports **511-513** are CAEs because the ports are capable of failing or generating errors and could be replaced after being properly diagnosed as a source of errors.

[0035] With reference now to **Figure 5B**, a table provides a topology map for the SAN shown in **Figure 5A** in accordance with a preferred embodiment of the present invention. The TM is represented as a two-dimensional table with both the left column and the top row containing the SAN elements, both CAE and CAN devices, connected to the switch, such as FC switch **501** in **Figure 5A**. The diagonal cells contain all the SPDIT/type information about the corresponding element and the switch port to which it is connected. The other cells contain the directional paths between the elements. For example, the table shows the direction path between CAE A and CAE C using the path between ports 3 and 1. Multiple paths are possible. The topology and registration discovery process are periodically repeated to ensure that the TM is current. The CAM will also register with any SAN elements providing Extended Link Services that can be used for PD.

[0036] With reference now to **Figure 6**, a diagram depicts a SAN Diagnostic Table for a SAN in accordance with a preferred embodiment of the present invention. The TM of a SAN is used to create a SAN Diagnostic Table (SDT) that is used for First Error Data Collection (FEDC) and Real-time Diagnostic Analysis (RDA). The SDT shown in **Figure 6** is similar to the TM shown in **Figure 5B** except that it contains an extra row for each switch/fabric element.

[0037] The diagonal SDT cells are used to hold the errors reported by the CAE corresponding to its row/column, including switch ports. Each point in a path, i.e. SDT cell, represents another SAN-connected element. Each cell contains the information collected in the TM so that specific product behaviors are known and proper diagnostic decisions can be made. For example, diagnostic queries may be presented, such as whether it is more likely that a storage device is causing a link error versus a Host Bus Attach (HBA) if out-of-band SCSI device errors accompany in-band HBA FC link errors.

[0038] The exemplary error information contained in **Figure 6** illustrates the utility of RDA using the SDT. Row 1 indicates that CAE A has reported an in-band FC link timeout. Row 3 indicates an out-of-band hardware controller error on CAE C. These two errors are related because they occurred in the same time frame, as shown by the timestamps associated with the error information. Row 5 indicates that an in-band FC link error has occurred, but given the stored timestamp, the error in row 5 is unrelated to the previous two. Therefore, the table depicts two separate problems: the first is related to a controller hardware failure in CAE C, and the second is a FC Link failure on CAE 2 in the FC Switch.

[0039] With reference now to **Figure 7**, a table depicts the weightings to be used in real-time diagnostic analysis for various errors in accordance with a preferred embodiment of the present invention. The RDA algorithms traverse error reporting elements of the SDT whenever an FEDC event occurs in order to determine the appropriate response. The RDA uses weighted decision analysis in order to isolate the failing component. Two broad categories are illustrated with H=Highest, M=Middle, L=Lowest weighting. The SDT traversal algorithms and error weightings are dynamic and would be changed to accommodate the complexity of the SAN topology and the nature of its connected elements.

[0040] The weighting table shown in **Figure 7** provides a simple illustration of the strong-to-weak weighting scale that applies to a typical SAN environment. If the SAN grows to just a few 16 port switches with its associated hosts and peripherals, the number of possible nodes that can report errors due to a single disk drive error or HBA timeout error can grow to a large number. Without global end-to-end RDA diagnostic capability, the task of isolating the failing component becomes hit-or-miss. In a multi-vendor SAN, it is common for multiple intermittent, recoverable, device errors, i.e. soft errors, to go unnoticed by the host. Eventually, the device may encounter an unrecoverable error, i.e. a hard error, that results in a system crash. The in-band and out-of-band mechanisms provided by the preferred embodiment would detect and report the recoverable errors as soon as they occur.

[0041] With reference now to **Figures 8A-8D**, flowcharts depict a process for a real-time diagnostic algorithm (RDA) for SAN end-to-end fault isolation of a single failing SAN element in accordance with a preferred embodiment of the present invention. The RDA uses two dynamic mechanisms to isolate faults:

1. Temporal Correlation Window (TCW) – The TCW is scalar value, i.e. time range, used to constrain fault isolation searching of the SDT in the time dimension so that the probability of misdiagnosis is minimized in the time dimension.

2. Spatial Correlation Path (SCP) – The SCP is a data structure that is used to constrain fault isolation searching in the spatial domain of the SDT so that only known system-to-subsystem associations are scrutinized and so that the probability of misdiagnosis is minimized in the spatial dimension. The SCP copies elements from the SDT during the RDA.

[0042] The goal of the RDA is to correlate all fault information received in time, location, and severity until the source of the fault is isolated with a high degree of certainty. This process terminates after a single reported fault or after a series of reported faults.

[0043] The general RDA for SAN end-to-end fault isolation of a single failing SAN element is described as follows. The process begins when a CAM initializes all connected paths in the CA into the SDT (step **801**). The SDT is initialized with all connected paths, i.e., paths A-->B, B-->C, etc. Only those paths that should be able to make connections are entered. These paths are established by the topology mapping, such as a TM similar to the TM shown in **Figure 5**. The SAN may not be fully connected in order to zone off certain connections that should not be able to connect. For example, certain hosts might be restricted to storing and retrieving data on particular storage devices. A system administrator may only allow NT hosts to store data on a particular device so that a mainframe does not have the ability to corrupt or destroy the NT data.

[0044] The process continues with a CAM initializing the TCW and SCP for the SAN (step **802**). The TCW is a time window and requires a time value, usually on the order from seconds to minutes. The SCP contains all sets of paths chosen from the SDT. These paths reflect known host-to-storage, host-to-host, and storage-to-storage associations that are established by the topology mapping. Again, it should be noted that a secondary CAM maintains a replica of the data structures and values that are stored in the primary CAM.

[0045] The CAM then receives a new error (step **803**) and processes the error using the RDA (step **804**). A determination is then made as to whether the RDA process is being terminated (step **805**), and if not, the process then loops back to step **803** to receive and process more errors. If so, then the process of initializing for SAN end-to-end fault isolation is complete.

[0046] Referring now to **Figure 8B**, a process depicts the processing of a new error, such as step **804** in **Figure 8A**, in more detail in accordance with a preferred embodiment. The process begins by receiving a new error (step **810**), and the SDT is updated to indicate the component reporting the error, the time the error occurred, and the severity (high, medium, low) of the error (step **811**). A determination is made as to whether the error is a high severity error (step **812**). If so, then this error is immediately reported as a fault that requires maintenance (step **813**). The SPDIT is then interrogated to determine if the reported error is associated with a specific part that should be replaced (step **814**). If not, then the processing of the high severity error is complete. If so, then the failing component is called out to be replaced (step **815**), and the processing of the high severity error is complete.

[0047] If the error is not a high severity error, then a determination is made as to whether the error is a medium or low severity error (step **816**). If so, then the low/medium severity error is processed (step **817**), and the error processing is complete.

[0048] If the error is neither a high severity error or a low/medium severity error, then the error severity is determined to be faulty and the error ignored (step **818**).

[0049] Referring now to **Figure 8C**, a process depicts the processing of a new low/medium severity error, such as step **817** in **Figure 8B**, in accordance with a preferred embodiment of the present invention, in more detail. The SCP is used to determine the paths that can be affected by the reported error. Each of the SDT cells for the elements in these paths, including the element reporting the new error, are interrogated in turn for previous occurrences of errors (step **820**), and it is determined if the occurrence of a previous error is spatially related to the current error (step **821**). The interrogation then uses the TCW in order to determine if the occurrence of a previous error is related to the current error in time as well as space (step **822**). If the previous errors are temporally and spatially related, then the errors are stored into the SCP (step **823**). After the interrogation is finished, the SCP contains the mapping of all errors on the appropriate paths in the SDT that occur within the time constraint.

[0050] The manner in which the data structure for the SCP is organized and used may vary depending upon system implementation. For example, the elements from the SDT may be copied into the SCP, and as errors are determined not to be related in space or in time, the elements may be deleted from the SCP.

[0051] The algorithm preferably makes an error correlation/severity assessment in order to isolate the location of the failing component. Referring now to **Figure 8D**, a flowchart depicts several possible cases for failing components associated with low/medium severity errors in accordance with a preferred embodiment of the present invention.

[0052] The process begins with a determination of whether all errors emanate from the current element which generated the newly received error (step **830**). If so, then a determination is made as to whether two or more errors are in the SCP (step **831**). If not, then the processing of the current error is complete. If so, then the current element is indicated to require maintenance (step **832**). The SPDIT is then interrogated to determine if the reported error is associated with a specific part that should be replaced (step **833**). If so, then the failing component is called out to be replaced (step **834**), and the newly received, low/medium severity error has been processed.

[0053] If all errors do not emanate from the current element, then a determination is made as to whether all (two or more) errors are contained in a single path (step **835**). In this case, any element in the path may be the root cause of the reported errors, and device hardware related errors take precedence over link related or timeout related errors. A determination is made as to whether the errors contain a device hardware error (step **836**).

If a device hardware error is found, in a manner similar to steps **832-834** described above, the associated element is indicated to require maintenance, the SPDIT is then interrogated to determine if the reported error is associated with a specific part that should be replaced, and if so, the failing component is called out to be replaced.

[0054] If the errors on the single path do not contain a device hardware error, then only link or timeout errors are being reported. This situation can lead to degradation in performance and eventual failure of the link. In this case, the algorithm looks for the element that is reporting the error first (step **837**), i.e., the first error takes precedence and the others are assumed to be related to the first occurrence. Once the element that is originating the chain of errors is found, in a manner similar to steps **832-834** described above, the associated element is indicated to require maintenance, the SPDIT is then interrogated to determine if the reported error is associated with a specific part that should be replaced, and if so, the failing component is called out to be replaced.

[0055] If two or more errors are not contained in a single path, then two or more errors are occurring on multiple paths. A determination is made as to whether there are any common elements on the paths of the multiple errors (step **838**), and if so, then this case requires isolating the common element(s) on these paths (step **839**) and performing an error correlation/severity assessment.

[0056] The common elements can either be SAN endpoint elements and/or SAN fabric elements. A determination is made as to whether a SAN endpoint or fabric element is the only common element (step **840**). If so, then in a manner similar to steps **832-834** described above, this common element is indicated as failing and maintenance is required. The SPDIT is then interrogated to determine if the reported error is associated with a specific part that should be replaced. If so, the failing component is called out to be replaced.

[0057] Otherwise, if a SAN endpoint or fabric element is not the only common element, then both a SAN endpoint and a SAN fabric element are common. This situation is now equivalent to the result from the determination in step **835**, and the process branches to step **836** for further processing.

[0058] If there are two or more errors that are not contained in a single path and there are no common elements on the paths of these errors, then each of the multiple errors are run through the real-time diagnostic algorithm (RDA) separately (step **841**). This rare but possible scenario may occur when more than one error has been received within the TCW and the errors originate from separately failing components. At this point, the error process may branch back to step **804** to process each error as if each error were a newly received error.

[0059] A SAN Diagnostic Table is created using the SAN topology, native Fibre Channel services, and vendor specific information. The present invention preferably supports both FC native in-band and host/device specific out-of-band status/error data collection for SAN problem determination. A real-time diagnostic algorithm may preferably then traverse the SAN Diagnostic Table to isolate a failing SAN component. The methodology is advantageous because it may be implemented on host operating environments such that special access to management terminals or device diagnostics are not, according to the preferred embodiment, required to isolate failing components. In addition, the methodology is platform-independent, and it supports both FC native in-band and host/device-specific out-of-band status/error data collection for SAN problem determination.

[0060] It is important to note that while the present invention has been described in the context of a fully functioning data processing system, those of ordinary skill in the art will appreciate that the processes of the present invention are capable of being distributed in the form of a computer readable medium of instructions and a variety of forms and that the present invention applies equally regardless of the particular type of signal bearing media actually used to carry out the distribution. Examples of computer readable media include recordable-type media such a floppy disc, a hard disk drive, a RAM, and CD-ROMs and transmission-type media such as digital and analog communications links.

Claims

1. A method for processing errors within a storage area network (SAN) (300), the method comprising the computer-implemented steps of:

   generating a SAN topology map;

   generating a SAN problem determination information table (SPDIT) (400) ; and

   generating a SAN diagnostic table (SDT) using the SAN topology map and the SPDIT.

2. The method of claim 1 wherein the SAN topology map comprises a table in which each row of the SAN topology table is uniquely mapped to a communication architecture element (CAE) (302, 304, 306, 311, 312, 313, 321, 322, 323, 324, 325) and each column of the SAN topology table is uniquely mapped to a CAE, wherein a CAE is a network-connected device that has successfully registered with a communications architecture manager (CAM) (302, 304, 308) via a network service protocol, wherein a CAM contains problem determination (PD) functionality for the SAN and maintains a SAN PD information table (SPDIT), and wherein the communication architecture (CA) comprises all network-connected elements capable of communicat-

ing information stored in the SPDIT.

3. The method of claim 2 wherein the SPDIT comprises at least one data record associated with each product or element on the CA.

4. The method of claim 3 wherein at least one data record associated with each product or element on the CA further comprises one or more data items selected from the group consisting of: product vendor information (401); product identifier information (402); information concerning a type of communication link supported by the product or element (403); and/or information concerning a type of error information to be reported by the product or element (404).

5. The method of claim 4 wherein the type of error information indicates whether the product or element supports Extended Link Services (ELS) Registered Link Incident Record (RLIR).

6. The method of claim 4 or 5 wherein the SDT stores information from the SAN topology map and errors received by the CAM from CAEs.

7. The method of any of claims 2 to 6, wherein the CA is managed by the CAM, the method further comprising:

receiving an error message at the communication architecture manager (CAM); and

processing the error message using a real-time diagnostic algorithm (RDA).

8. The method of any of claims 2 to 7 wherein a network supporting the CA comprises in-band Fibre Channel communication links and out-of-band communication links.

9. The method of any of claims 2 to 8 wherein the SAN comprises:

a plurality of storage devices connected to the network; and

a plurality of host computers connected to the network, wherein at least one of the plurality of host computers comprises a CAM.

10. The method of claim 7 further comprising:

analyzing the received error message using a temporal correlation window (TCW) value to temporally constrain fault isolation determination while searching for temporally-related error messages previously received by the CAM and stored within the SDT; and

analyzing the received error message using a spatial correlation path data structure (SCP) to spatially constrain fault isolation determination while searching for spatially-related error messages previously received by the CAM and stored within the SDT.

11. The method of claim 10 further comprising:

analyzing the received error message using error severity weightings according to a type of error indicated by the received error message.

12. A data processing system for communicating error information in a storage area network (SAN) (300), the data processing system comprising:

a network comprising in-band Fibre Channel communication links and out-of-band communication links, wherein the network supports a communications architecture (CA);

a plurality of storage devices connected to the network;

a plurality of host computers (302, 304, 306) connected to the network, wherein at least one of the plurality of host computers comprises a communications architecture manager (CAM) (302, 304) containing problem determination (PD) functionality, wherein a CAM maintains a SAN PD information table (SPDIT) (400), and wherein the CA comprises all network-connected elements capable of communicating information stored in the SPDIT.

13. The data processing system of claim 12 further comprising:

a plurality of CAMS, wherein the CA comprises a primary CAM (304) and one or more secondary CAMS (308), wherein a secondary CAM operates redundantly for a primary CAM.

14. The data processing system of claim 12 or 13 wherein the CA further comprises or more CA elements (CAEs) and one or more CA non-participants (CANs) (325), wherein a CAE is a network-connected device that has successfully registered with a CAM via a network service protocol, and wherein a CAN is a network-connected device that has not registered with a CAM yet known to be present via a SAN topology discovery process.

15. The data processing system of claim 12, 13 or 14 wherein the wherein the in-band Fibre Channel communication links and the out-of-band communication links are provided by a single, physical

communication link.

16. A data processing system for processing errors within a storage area network (SAN), the data processing system comprising:

first generating means for generating a SAN topology map;

second generating means for generating a SAN problem determination information table (SPDIT); and

third generating means for generating a SAN diagnostic table (SDT) using the SAN topology map and the SPDIT.

17. A computer program comprising program code for controlling a data processing system to perform operations implementing a method for processing errors within a storage area network (SAN) (300), the program code comprising:

first instructions for generating a SAN topology map;
second instructions for generating a SAN problem determination information table (SPDIT) (406); and
third instructions for generating a SAN diagnostic table (SDT) using the SAN topology map and the SPDIT.

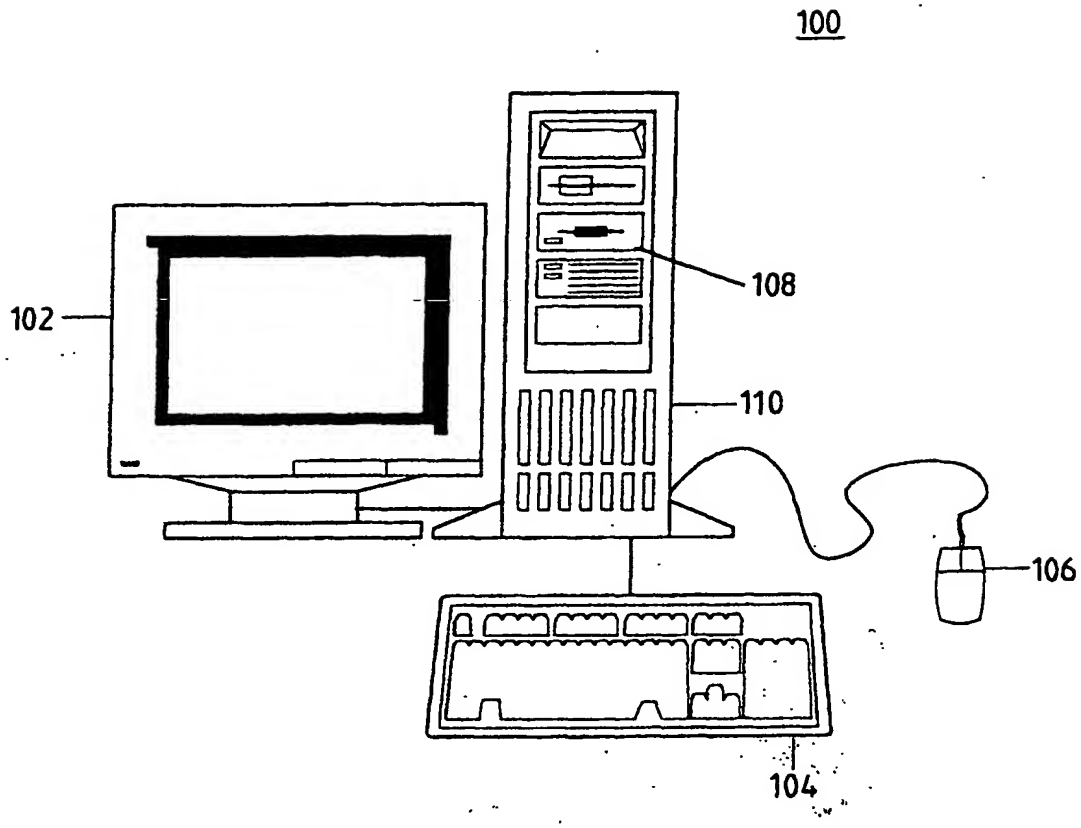

FIG. 1

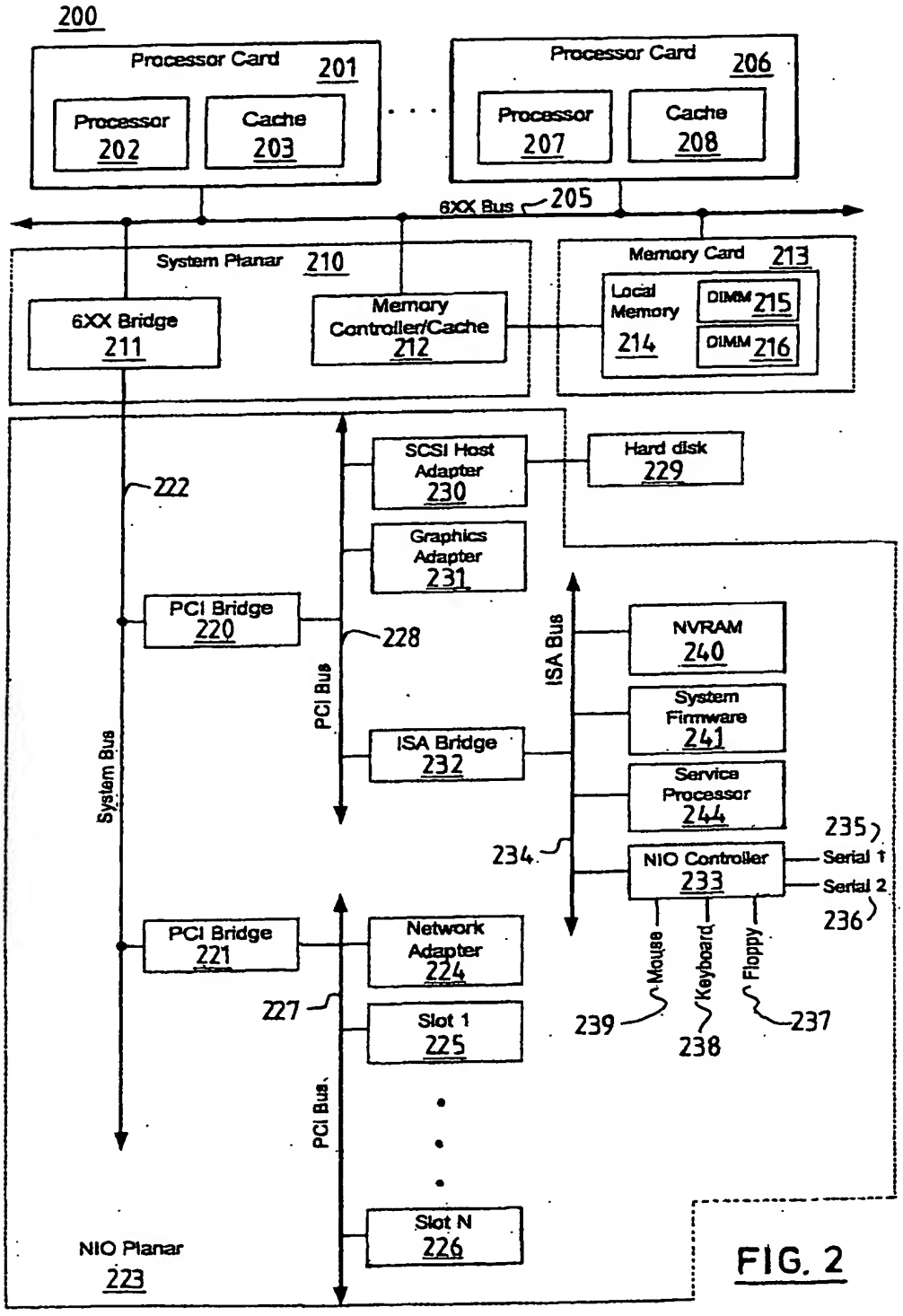
200
Processor Card 201
Processor 202
Cache 203
Processor Card 206
Processor 207
Cache 208
6XX Bus 205
System Planar 210
6XX Bridge 211
Memory Controller/Cache 212
Memory Card 213
Local Memory 214
DIMM 215
DIMM 216
222
SCSI Host Adapter 230
Hard disk 229
Graphics Adapter 231
PCI Bridge 220
228
PCI Bus
ISA Bus
NVRAM 240
System Firmware 241
ISA Bridge 232
Service Processor 244
235
234
NIO Controller 233
Serial 1
Serial 2
236
System Bus
PCI Bridge 221
Network Adapter 224
227
Slot 1 225
Mouse
Keyboard
Floppy
239
238
237
PCI Bus
Slot N 226
NIO Planar 223

FIG. 2

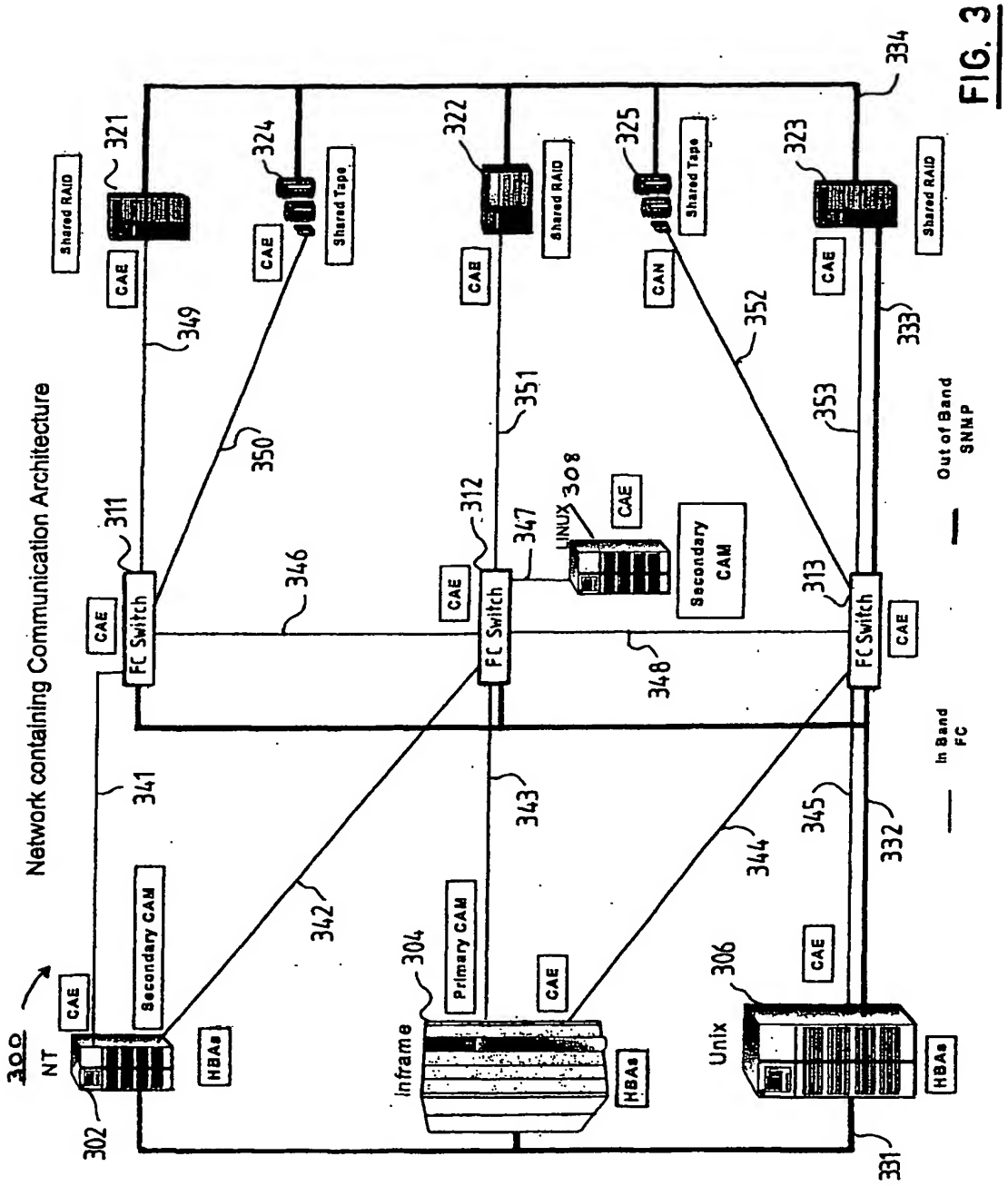
Network containing Communication Architecture
300
NT
Inframe
Unix
Shared RAID
Shared Tape
Secondary CAM
Primary CAM
FC Switch
CAE
HBAs
LINUX
CAN
In Band FC
Out of Band SNMP
302
304
306
308
311
312
313
321
322
323
324
325
331
332
333
334
341
342
343
344
345
346
347
348
349
350
351
352
353

FIG. 3

SAN Problem Determination Information Table (SPDIT)

400

Vendor 401	Product 402	Info Type 403	Description 404
FC Switch Corp.	16 F_Port FC Switch, Model 123 ProdID=1	In-Band: FC Link, FC Extended Link Services: Registered Link Incident Record	ELS RLIR Compatible
...	...	...	...
FC Storage Corp.	4-Port FC RAID Model 234 ProdID=2	Out-of-band: SNMP MIB In-band: FC Link Extended Link RLIR	MIB=FCSTORINC ELS RLIR Compatible

FIG. 4

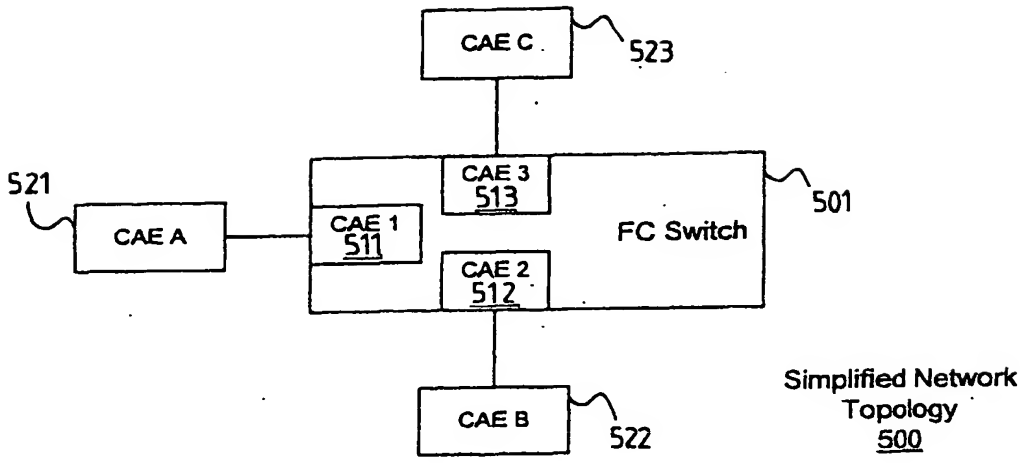

FIG. 5A

	CAE A	CAE B	CAE C
CAE A	SPDIT/TYPE Information Switch Port CAE 1	(Path A–>B) = (1–>2) Path not used.	(Path A–>C) = (1–>3)
CAE B	(Path B–>A) = (2–>1) Path not used.	SPDIT/TYPE Information Switch Port CAE 2	(Path B–>C) = (2–>3)
CAE C	(Path C–>A) = (3–>1)	(Path C–>B) = (3–>2)	SPIDT/TYPE Information Switch Port CAE 3

Topology Map for SAN

FIG. 5B

		CAE A	CAE B	CAE C
Row 1	CAE A	Posted Error for CAE A In-Band Link Timeout 1/1/99@2:00.55		(Path A-->C) = (1-->3)
Row 2	CAE B		Posted Error for CAE B	(Path B-->C) = (2-->3)
Row 3	CAE C	(Path C-->A) = (3-->1)	(Path C-->B) = (2-->3)	Posted Error for CAE C Out-of-Band Controller Fault 1/1/99@2:00.01
Row 4	Switch CAE 1	Posted Error for CAE 1		
Row 5	Switch CAE 2		Posted Error for CAE 2 In-Band Link Error 1/2/99@4:55.09	
Row 6	Switch CAE 3			Posted Error for CAE 3

SAN Diagnostic Table (SDT)

FIG. 6

Weightings for Device Errors	Weightings for Link Errors
Controller Hardware Failure = H	Link Hardware Non-Operational (Switch/N/L) = H
Device Hardware Failure = H	Link Hardware Failure Recoverable = M
Dev Hardware Error Recoverable = H	Link Timeout Error = L
Device Timeout = M	Link Hardware Degraded = H
Controller Timeout = M	Power Hardware Fault = H
Power Hardware Fault = H	

Error Weightings for Real-Time Diagnostic Analysis

FIG. 7

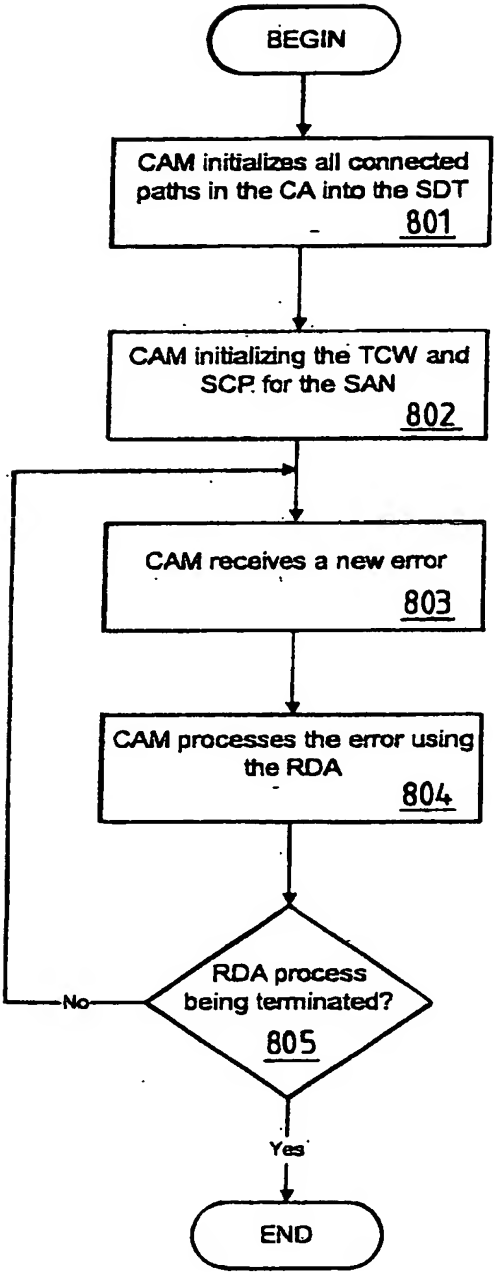
BEGIN
CAM initializes all connected paths in the CA into the SDT
801
CAM initializing the TCW and SCP for the SAN
802
CAM receives a new error
803
CAM processes the error using the RDA
804
RDA process being terminated?
805
No
Yes
END

FIG. 8A

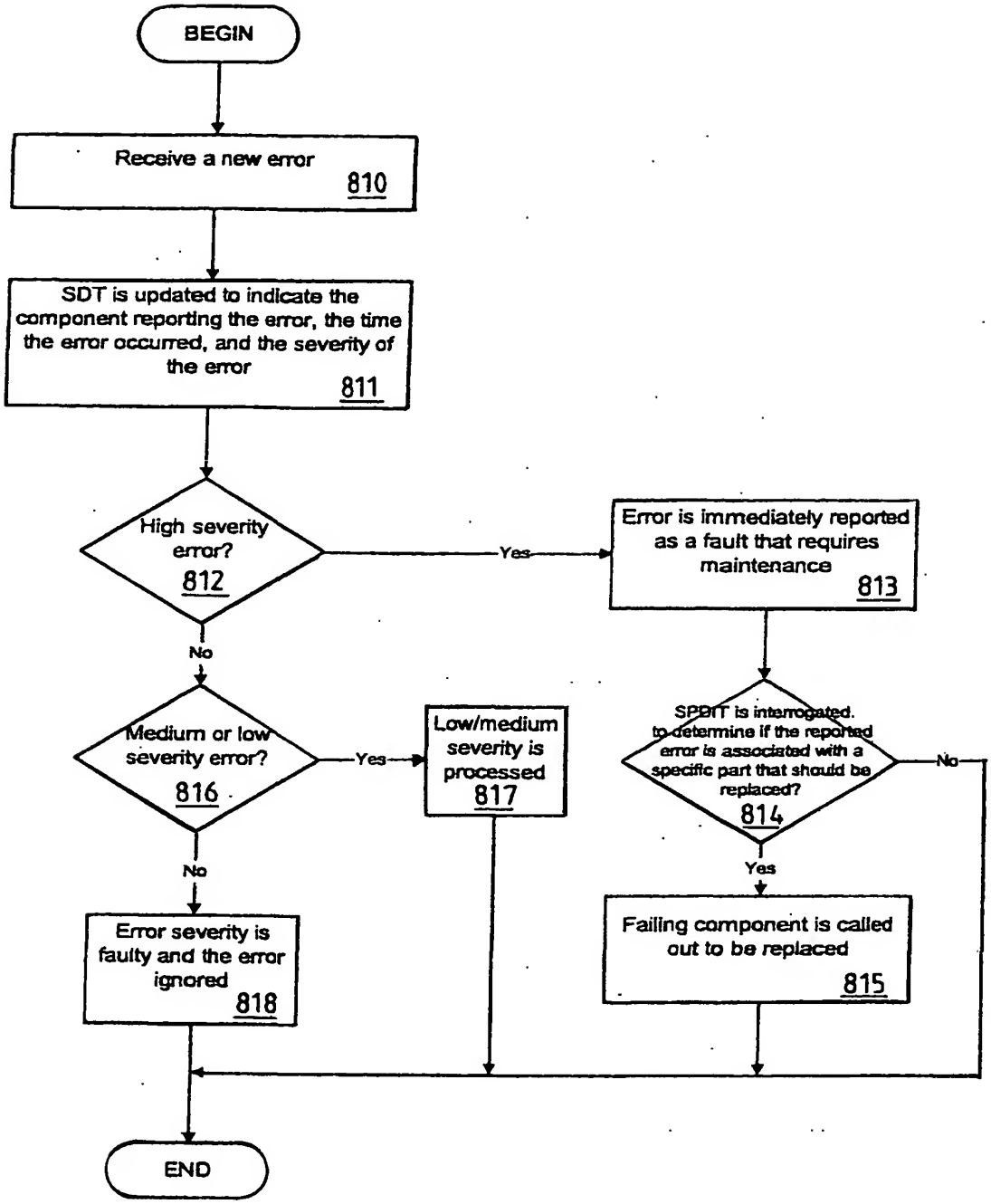
BEGIN
Receive a new error
810
SDT is updated to indicate the component reporting the error, the time the error occurred, and the severity of the error
811
High severity error?
812
Yes
Error is immediately reported as a fault that requires maintenance
813
No
Medium or low severity error?
816
Yes
Low/medium severity is processed
817
SPDIT is interrogated, to determine if the reported error is associated with a specific part that should be replaced?
814
No
Yes
Failing component is called out to be replaced
815
No
Error severity is faulty and the error ignored
818
END

FIG. 8B

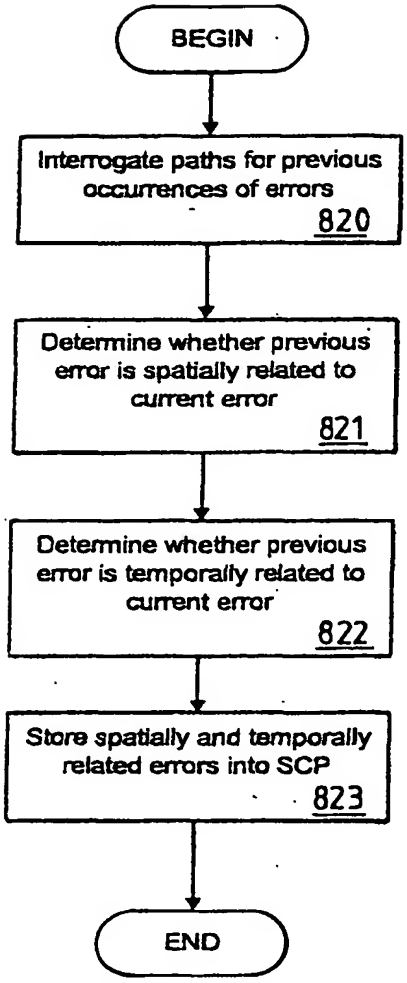
BEGIN
Interrogate paths for previous occurrences of errors
820
Determine whether previous error is spatially related to current error
821
Determine whether previous error is temporally related to current error
822
Store spatially and temporally related errors into SCP
823
END

FIG. 8C

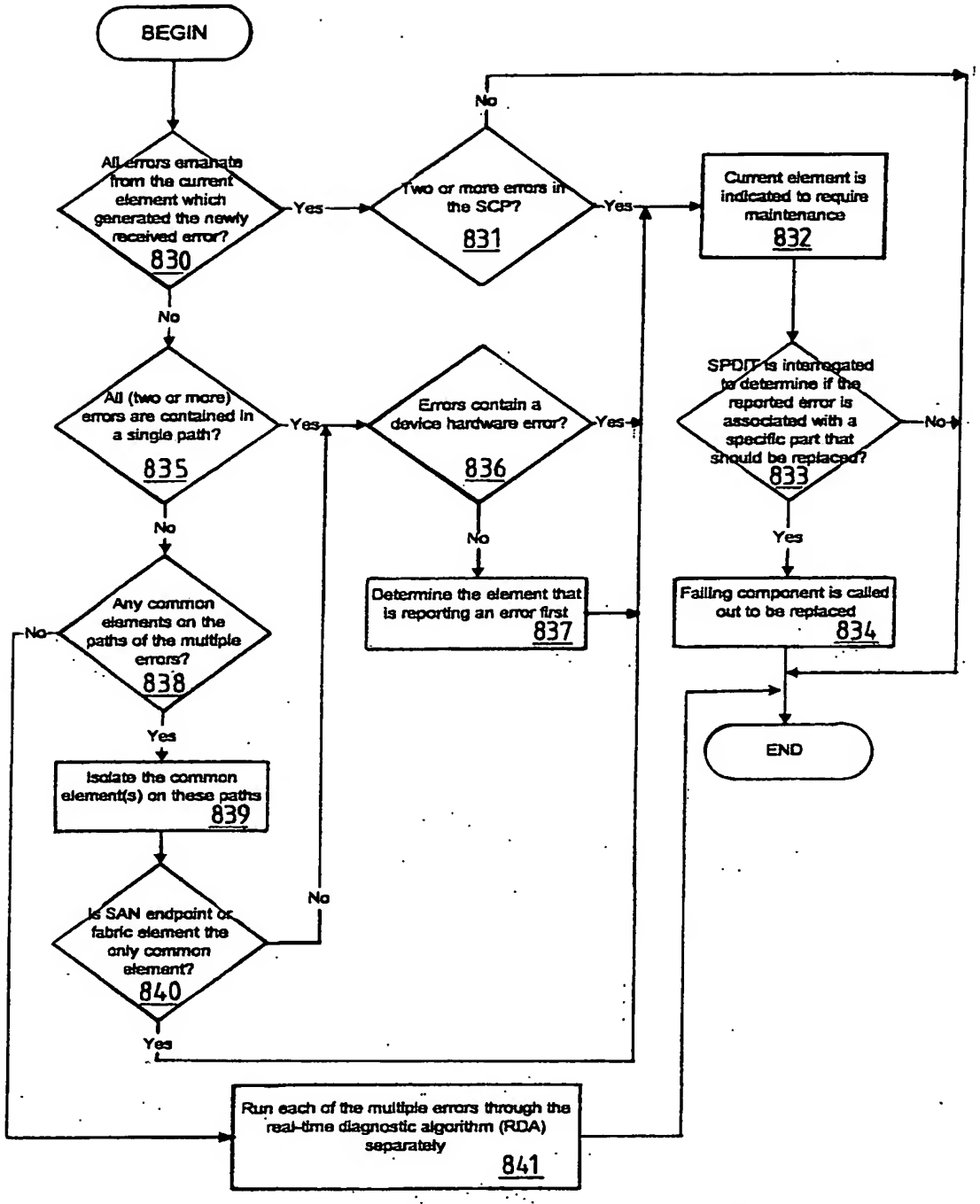
BEGIN
All errors emanate from the current element which generated the newly received error?
830
Yes
Two or more errors in the SCP?
831
Yes
Current element is indicated to require maintenance
832
No
No
All (two or more) errors are contained in a single path?
835
Yes
Errors contain a device hardware error?
836
Yes
SPDT is interrogated to determine if the reported error is associated with a specific part that should be replaced?
833
No
No
Yes
No
Any common elements on the paths of the multiple errors?
838
No
Determine the element that is reporting an error first
837
Failing component is called out to be replaced
834
Yes
END
Isolate the common element(s) on these paths
839
Is SAN endpoint or fabric element the only common element?
840
No
Yes
Run each of the multiple errors through the real-time diagnostic algorithm (RDA) separately
841

FIG. 8D